AD_____

AWARD NUMBER DAMD17-94-J-4406

TITLE: Statistical Genetics Methods for Localizing Multiple Breast Cancer Genes

PRINCIPAL INVESTIGATOR: Jurg Ott, Ph.D.

CONTRACTING ORGANIZATION: Columbia University
New York, New York 10032

REPORT DATE: September 1998

TYPE OF REPORT: Final

PREPARED FOR: Commanding General
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1998	Final (1 Sep 94 - 31 Aug 98)

4. TITLE AND SUBTITLE
Statistical Genetics Methods for Localizing Multiple Breast Cancer Genes

5. FUNDING NUMBERS
DAMD17-94-J-4406

6. AUTHOR(S)
Jurg Ott, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Columbia University
New York, New York 10032

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The purpose of this work was to develop statistical methods for improving current statistical analysis methods of genetic linkage and linkage disequilibrium. A method to calculate support (confidence) intervals for genetic risks was extended to include uncertainties in penetrance values. For a number of variables measured on a trait, a method, principal components of heritability, was developed that combines these variables in such a way that the resulting linear combination has highest heritability. For diseases of late onset, in which often only affected sib pairs are available, a method was developed to detect pairs that are not true siblings (they may be half-sibs or be unrelated) and eliminate them from the analysis. For the analysis of linkage disequilibrium, rates of true and apparent errors were investigated analytically and by computer simulation. This showed that in extreme situations, true error rates may be four times higher than the error rates detected as mendelian inconsistencies.

14. SUBJECT TERMS
Breast Cancer

15. NUMBER OF PAGES
62

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

_____ Where copyrighted material is quoted, permission has been obtained to use such material.

_____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

_____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

_____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

_____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

_____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

_____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

_____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Jwp on 9/29/98

PI - Signature Date

TABLE OF CONTENTS

INTRODUCTION

The classical approach to elucidating the etiology of diseases has been to investigate the pathways that are involved in the disease process. For genetic traits, a reverse approach has become more and more important. It is called positional cloning and does not require initially knowing the biochemical pathway leading to disease. Instead, a disease gene is identified by genetic means, which, in turn, allows pinpointing the process underlying the trait.

The current approach to positional cloning consists essentially of two steps: (1) Localizing a disease gene on the human gene map by way of genetic linkage analysis, and, based on this result, (2) identifying the disease gene by molecular genetics methods (cloning). The work described here focuses on the first step by developing and implementing statistical methods to more accurately localize disease genes. For breast cancer, genetic linkage analysis has already been successful in localizing two disease genes, BRCA1 and BRCA2, which have subsequently been cloned.

For many years, genetic linkage analysis had to be carried out essentially by pencil and paper. It was close to 25 years ago that the PI developed the first generally available computer program for genetic linkage analysis in human families (Ott 1974), which soon was adopted as a standard worldwide. However, with ongoing developments in molecular genetics, there must be a parallel development in statistical analysis methods. For example, with today's large number of genetic markers available, even the newer approaches to linkage analysis have become cumbersome to carry out. In the linkage analysis of the breast cancer consortium (Easton et al. 1993), "a full multipoint analysis including all [six] markers and BRCA1 was not feasible." This is just one example of an area in need of improvement.

The purpose of the work described here is to improve current analysis methods to make linkage analysis methods more accurate. Four specific aims were formulated as follows:

1. To extend a previously developed method of determining support intervals for *genetic risks* to the specific situation of the BRCA1 gene, that is, to allow for uncertainties in the age at onset parameters.
2. To further develop the *polylocus method of linkage analysis* between a disease and a set of closely linked markers and to extend it to the case of a disease gene with flanking markers.
3. To develop a computer simulation method for *approximating the maximum lod score* in situations where exact calculation of the lod score is prohibitively time consuming.
4. To improve the currently available methods for assessing linkage through allelic association (*linkage disequilibrium*).

Specific aim 1 refers to the situation that a breast cancer gene is known and that genetic risks are to be calculated for relatives of affected women. Aims 2 and 3 focus on the complexities of lod score calculations in genetic linkage analysis and propose approximations to make the search for additional breast cancer genes more efficient.

Specific aim 4 proposes improvements to finding gene loci through linkage disequilibrium. The concerted effect of these approaches is expected to greatly enhance the prospects of finding additional breast cancer genes.

BODY

The methods employed in this work are all of a statistical nature. In fact, it is methods development and its applications that are the object of the work carried out. The material covered is ordered by specific aim. Positive and negative findings are outlined in each section. Two of the specific aims (numbers 2 and 3) had to be modified because, as originally proposed, they turned out to be superseded by recent developments.

1. Support intervals for genetic risks

In a mendelian trait, the genetic risk is the conditional probability that an individual has the genetic susceptible genotype given both phenotype and genotype information for all available pedigree members. Genetic risks may be based on the pedigree likelihood as originally proposed by Elston and Stewart (1971). In addition to such genotype risks, a phenotype risk may be defined as the conditional probability of developing the trait. With incomplete penetrance and absence of phenocopies, the phenotype risk is smaller than the corresponding genotype risk. Generally, however, phenocopies as well as genetic cases contribute to the phenotype risk.

The precision of risk estimates is dependent on the accuracy of the parameters used in their evaluation. Usually risks are computed under the assumption that genetic parameters are known without error. Uncertainty in the accuracy of parameter estimates renders uncertainty in the risk. Therefore, in order to evaluate the accuracy of a risk, it is critical to calculate either a confidence or support interval for the risk (Weeks and Ott 1989; Ott, 1991).

Previously, we described a method to construct support intervals (SIs) for genetic risks working in a maximum likelihood framework (Leal and Ott 1994). Briefly, the method allows for parameters (disease allele frequency, recombination fraction, etc.) to vary in their support intervals. For each combination of parameter values so obtained, a risk is calculated whose associated log likelihood is equal to the log likelihood at the given parameter values. All those risk values with a log likelihood within *m* units of the maximum log likelihood form the risk support interval. Under this grant, this method was extended to allow for variability of genotype-specific *penetrances* when the age at disease onset is normally distributed. As an empirical example, the SIs for the phenotype and genotype risk have been calculated for a member of a breast-ovarian cancer kindred using two markers (D17S250 and D17S588) which are linked to the BRCA1 locus.

Specifically, genotype-specific penetrance probabilities are incorporated in the calculation of SIs for the genotype and phenotype risk in the following manner: Approximate *m*-unit SIs are constructed around the mean age of disease onset, μ, and lifetime penetrances, λ, each for disease gene carriers and noncarriers. The calculation of maximum and joint log likelihoods for all parameters is carried out as previously described (Leal and Ott 1994), except that here, the estimates, μ, for age at disease onset are taken to

follow a normal distribution while all other parameter estimates are binomially distributed. In the likelihood calculations, a penetrance probability is the genotype-specific cumulative risk for unaffected and affected individuals when age of onset is unknown, and the genotype-specific density for affected individuals when age of onset is known.

Each parameter is varied within its SI. When the joint log likelihood for a set of parameter values falls within *m* units of maximum log likelihood, the genotype-specific penetrance probabilities are calculated for each liability class and the risk is calculated with the aid of MLINK (Lathrop and Lalouel 1984). The phenotype risk is also computed using a specific cumulative penetrance liability class. The highest and lowest (genotype and phenotype) risks so obtained are taken to be the endpoints of the (genotype and phenotype) risk SI. These changes have been implemented in the RISKSI program and described in a paper (Leal and Ott 1995).

As an empirical *example*, 2-unit SIs for the phenotype and genotype risk were calculated for individual 405, an unaffected 52 year old female who is a member of the breast-ovarian cancer kindred CRC101 (Smith et al. 1993), given her current age. Technical details may be found in the paper (Leal and Ott 1995). The resulting SI for the genotype risk that she carries the BRCA1 susceptibility allele ranges from 0 through 14.5%, and the SI for her phenotype risk is between 5.9% and 19.4%. The point estimates for the genotype and phenotype risk are 2.1% and 8.4%, respectively. Clearly, it is important for a counselor to know how accurate a risk estimate is and how far the true risk may deviate from the estimated risk.

2. Accuracy of linkage analysis: phenotype

In the original application, an approach called polylocus linkage analysis method was quoted. It was proposed to extend that method to more realistic situations because it seemed to offer ways of handling multiple highly polymorphic markers jointly, with is not possible with the LINKAGE programs. However, in the meantime, analysis methods have been introduced that can do just that (Kruglyak et al. 1996, implemented in the GENEHUNTER program). Therefore, another aspect of linkage analysis was chosen that was in need of improvement.

For many traits, it is unclear what the relevant phenotype is that is under the control of an underlying gene. Whereas there exist unequivocal diagnostic schemes for diseases, the genetically relevant disease definition is often unclear. Many variables are often measured. Either each of them is subjected so a separate analysis, or the items may be used together as a multivariate outcome in a single analysis (Amos et al. 1990). With a large number of items, however, the power of a multivariate analysis to detect linkage can be substantially lower than the power of an analysis applied to a genetically relevant scale. This is because the increased degrees of freedom for the reference distribution of a multivariate analysis can result in too stringent a criterion for statistical significance. One approach to choosing a scale is to subject the available items to a principal components analysis. The first few principle components are then candidates for linkage analyses. An intermediate step may be to estimate the heritability of the first several principal components, and to use the components with highest heritability. See Hasstedt et al (1994) for an example concerning sodium transport and Livshits (1995) for an example concerning body size and shape. However, principle components depend on the

variance covariance matrix of the data pooled from all pedigrees, and thus do not reflect family structure. By focusing only on scales obtained as principal components, other scales with higher heritability may be overlooked.

Here, an approach was developed in which the available items are combined into scales, not on the basis of their variance-covariance structure, but instead, on the basis of *heritability*. The first scale has highest heritability, the second scale has highest heritability among all scales uncorrelated with the first, the third has highest heritability among those uncorrelated with the first and second, and so on. Heritability is the ratio of the variances of family specific components and individual specific components of variance. Thus, a combination of the variance-covariance structure of both the between family and within family variance components are used to compute the scales.

The theory underlying this approach is quite technical and somewhat cumbersome to describe. For the statistical derivation, see Ott and Rabinowitz (1999), a copy of which is provided in the appendix.

Simulation experiments show that this approach has clear advantages when some of the items are under genetic control while others are not (see third simulation experiment in Ott and Rabinowitz 1999). In that experiment, five items were considered, of which two were genetically determined (at a disease locus) while three were not. Power for detecting linkage (significance level, $\alpha = 0.001$) of the conventional principal components approach was 0.011 and that of the principal components of heritability approach was 0.744.

3. Handling errors in linkage analysis

As originally proposed, a method to *approximate* lod scores (rather than to calculate them exactly) was developed on the basis of computer simulation in a specific pedigree. It turned out, however, that the number of replicates required for reasonable accuracy was so large that the method was not any faster than exact calculation of lod scores. In addition, such methods were successfully being developed by Markov Chain Monte Carlo methods (Heath 1997) so that there was no point in pursuing this any further.

Therefore, the originally proposed method was modified in favor of a new approach to optimize linkage analysis. This approach also involves approximation in the sense that an estimation procedure was developed to screen for sib pairs with errors, which are then removed from analysis. Computer simulation of the method demonstrates that its application leads to increased power in linkage analysis.

Currently, two breast cancer genes, BRCA1 and BRCA2, are known, yet many breast cancer families show absence of linkage to either of these genes. Thus, investigators are searching for additional genes responsible for familial breast cancer. These genes are expected to be associated with late onset breast cancer while BRCA1 occurs primarily in early onset disease. One of the problems with linkage analysis of late onset disease is that parents may be unavailable. Thus, in affected sib pair analysis (a widely used nonparametric linkage analysis technique), there is no way that errors can be detected through mendelian inconsistencies: The two sibs share at most four different alleles, which is the same number as there are alleles among two parents. Laboratory errors (sample swaps, allele misreading, etc.), unrecognized adoption, and other errors are

generally recognized through the occurrence of mendelian inconsistencies but this is not possible for two affected sibs with their parents missing. Such errors typically occur with frequencies of around 1% and have the consequence that the purported siblings may not be sibs but, for example, half-sibs or unrelated individuals. In linkage analysis, occurrence of errors greatly reduces informativeness (Buetow 1991). To offset this potential loss of information, a method was developed and implemented in a computer program to screen for non-sibs by statistical means and to remove them from the analysis. As shown below, application of this method greatly increases power of affected sib pair linkage analysis.

As a consequence of the mendelian laws of inheritance, the genotypes of two relatives are similar to each other. Consequently, by establishing whether or not the genotypes of two individuals are correlated and to what degree, it should be possible to determine whether these individuals are related or not, and perhaps what the degree of relationship is. On the basis of the genotypes for a set of unlinked marker loci, Thompson (1975, 1991) developed appropriate statistical theory to estimate the relationship between two individuals. To apply her approach to this project, her theory was extended to allow for unlinked *and linked* markers. In addition, Bayesian methods are applied that incorporate the typically known prior error probabilities. For details on the theory of relationship estimation, see the paper resulting from this work (Göring and Ott 1996).

A simple outline of the main steps of the new method is as follows: Based on the genotypes at all marker loci available in an affected sib pair study, for each stated sib pair the posterior probability is computed that the two individuals are siblings. When this probability falls below a certain threshold then the pair of individuals is discarded from the study (because they are assumed not to be siblings). The remaining stated sib pairs are then most likely pairs of true siblings and are analyzed in one of the usual affected sib pair approaches. The threshold was chosen on the basis of a decision rule that maximizes power if in fact linkage of a recessive trait to a fully informative marker exists. Computer simulation shows that this decision rule is conservative, that is, very few true siblings tend to be discarded from a study.

Even though the new method reduces the number of "sib" pairs available for analysis, it results in a dramatic increase of power because the sib pairs remaining in the analysis after application of the Bayesian relationship estimation are with high probability real sibs. For example, consider the following case (case *a* in table 5 of Göring and Ott 1996): 400 sib pairs without parents are available for study. They tend to consist of 98% true sibs, 1% half-sibs, and 1% pairs of unrelated individuals. 100 marker loci (70% heterozygosity each) are typed on each sib pair. The disease was taken to be a complex trait (multifactorial threshold trait) with population prevalence of 5% and heritability (on the liability scale) of 50%. The recombination fraction between the major disease locus and a marker locus was assumed to be 1%. The result of an affected sib pair analysis is considered significant when the empirical significance level is at most 0.0001, which approximately corresponds to a maximum lod score of 3. Under these conditions, power to detect linkage is 0.39 when all stated sib pairs are used. It increases to 0.51 when the new method is applied prior to the affected sib pair analysis. This value of 0.51 is also the power when the known non-sibs are removed prior to affected sib pair analysis, that is, the new method removes non-sibs with high confidence and has little tendency to remove true sibs.

As a practical application of linkage analysis to breast cancer families, a collaboration with researchers at Columbia University resulted in a paper on Cowden's syndrome and BRCA1 (Tsou et al. 1997). The role of the PI in that paper was restricted to carrying out an appropriate linkage analysis.

4. Linkage disequilibrium

Linkage analysis is one of the methods to localize disease genes. Another method consists of investigating whether in a set of affected individuals, marker alleles have a different frequency distribution than in a set of control individuals. If so, this is interpreted as evidence that disease alleles and marker alleles are associated due to proximity of disease and marker loci (linkage disequilibrium, LD). Two approaches to LD analysis were considered.

4.1. Modification of EH program

As proposed originally, a previously existing computer program, EH, was modified to allow for LD between a mendelian locus and a marker locus by modeling the penetrance structure of the mendelian trait. Thus, this approach is expected to be more powerful than simply looking for a difference in allele frequencies between affected and control individuals. The technical description of this approach may be found in the *Methodology* section of Ott (1998).

4.2. Investigation of errors in family trios

For the investigatin of LD, sampling designs have been developed ("haplotype relative risk" designs) in which the sampling unit is an affected offspring and his or her parents. Marker alleles transmitted to the offspring are contrasted to those not transmitted, where the latter form a contrived control sample (Falk et al. 1987, Ott 1989, Thomson 1995). The data of interest are genotypes at a marker for a family trio, namely a father, mother, and a child.

One of the aims in developing marker genes of a new type (SNPs, Single Nucleotide Polymorphisms) is the detection of disease loci by disequilibrium analysis. These SNPs have two alleles at each locus.

We consider the following problem: suppose that, for a random sample of families in which father, mother and child have been typed at a SNP locus, pedigree errors (any changes from one allele to another, e.g. due to sample swaps, genotyping error, etc.) occur randomly and independently with some fixed probability α. We say that an error in a trio is *undetectable* if the genotypes of the trio still display mendelian consistency. The question addressed is: what is the probability that errors in our sample of family trios go undetected? Let us label this quantity by β. We investigate the question for a single SNP locus with varying degrees of polymorphism. Once we calculate β, we can also calculate the detection rate, $1 - \beta$, and the apparent error rate, $\alpha(1 - \beta)$, with which an error is observed through a mendelian inconsistency.

The relevant methological approach and calculations are given in the paper provided in the appendix (Gordon et al. 1999). For a range of situations (see table 4 in

Gordon et al. 1999), if we use only Mendel's laws to check whether or not a pedigree error has occurred, then the true error rate is roughly 3.3 to 4 times the magnitude of the apparent error rate.

As a comparison to the SNPs scenario, we performed simulations in which we generated 100,000 genotype trios for a marker with ten equally frequent alleles. For various induced rates of genotyping error, the apparent error rate was recorded. In the case of a marker with 10 equally frequent alleles, the true error rate is approximately 1.3 to 1.7 times the apparent error rate.

These results demonstrate the vulnerability of the currently favored approach to investigating LD with family trios on the basis of SNP markers.

CONCLUSIONS

The work carried out under this research grant generally provides researchers with methods for improving the accuracy of linkage and disequilibrium analysis.

REFERENCES

Papers acknowledging this grant are not listed in this section but only in the bibliography section, below.

Amos CI, Elston RC, Bonney GE, Keats BJB, Berenson GS (1990) A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. *Am J Hum Genet* 47:247-254

Buetow KH (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am J Hum Genet* 49:985-994

Elston RC, Stewart J (1971): A general model for the analysis of pedigree data. *Hum Hered* 21:523-542

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 51:227-233

Hasstedt SJ, Hunt SC, Wu LL, Williams RR (1994) Evidence for multiple genes determining sodium transport. *Genet Epidemiol* 11:553-568

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* 61:748-760

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am J Hum Genet* 58:1347-1363

Lathrop GM, Lalouel JM (1984): Easy calculations of lod scores and genetic risks on small computers. *Am J Hum Genet* 36:460-465

Leal SM, Ott J (1994): A likelihood approach to calculating risk support intervals. *Am J Hum Genet* 54:913-917

Livshits G, Otremski I, Kobyliansky E (1995) Genetics of human body size and shape: complex segregation analysis. *Ann Hum Biol* 22:13-27

Ott J (1989) Statistical properties of the haplotype relative risk. *Genet Epidemiol* 6:127-130

Ott J (1991) *Analysis of human genetic linkage*. Baltimore: Johns Hopkins University Press

Smith SA, Easton DF, Ford D, Peto J, Anderson K, Averill M, Stratton M, Ponder M, Pye C, Ponder BJA (1993) Genetic heterogeneity and localization of a familial breast-ovarian cancer gene on chromosome 17q12-q21. *Am J Hum Genet* 52:767-776

Thomson G (1995) Mapping disease genes: family-based association studies. *Am J Hum Genet* 57:487-498

Weeks DE, Ott J (1989) Risk calculations under heterogeneity. *Am J Hum Genet* 45:819-821

APPENDIX

Copies of the following three publications are enclosed in the appendix: Gordon et al. (1999), Ott (1998), Ott and Rabinowitz (1999).

BIBLIOGRAPHY OF ALL PUBLICATIONS AND MEETING ABSTRACTS

All publications listed in this section acknowledge support by this research grant.

Gordon D, Heath SC, Ott J (1999) True pedigree errors more frequen than apparent errors for single nucleotide polymorphisms. *Hum Hered* (in press) (a copy is provided in the Appendix).

Göring HH, Ott J (1997) Relationship estimation in affected sib pair analysis of late-onset diseases. *Eur J Hum Genet* **5**, 69-77

Leal SM, Ott J (1995) Variability of genotype-specific penetrance probabilities in the calculation of risk support intervals. *Genet Epidemiol* **12**, 859-862

Ott J (1997) Methods for gene mapping and calculating risks in breast cancer. *Era of Hope*, Proceedings of The Department of Defense Breast Cancer Research Program Meeting, vol. I. Washington, DC, 1997.

Ott J (1998) User's guide to the EH program. Electronic document published by the Laboratory of Statistical Genetics, Rockefeller University, New York. Accessible as http://linkage.rockefeller.edu/ott/eh.htm (a copy is provided in the Appendix).

Ott J, Rabinowitz D (1999) A principal components approach based on heritability for combining phenotype information. *Hum Hered* (in press) (a copy is provided in the Appendix).

Tsou HC, Teng DH, Ping XL, Brancolini V, Davis T, Hu R, Xie XX, Gruener AC, Schrager CA, Christiano AM, Eng C, Steck P, Ott J, Tavtigian SV, Peacocke M (1997) The role of MMAC1 mutations in early-onset breast cancer: causative in association with Cowden syndrome and excluded in BRCA1-negative cases. *Am J Hum Genet* **61**, 1036-1043

LIST OF SALARIED PERSONNEL

Harald Göring, M.S., Department of Genetics and Development, Columbia University, New York

Jurg Ott, Ph.D., Department of Genetics and Development, Columbia University, New York

Human Heredity (In Press)

True Pedigree Errors More Frequent Than Apparent Errors for Single Nucleotide Polymorphisms

Derek Gordon, Simon C. Heath, and Jurg Ott

Laboratory of Statistical Genetics

Rockefeller University, Box 192

1230 York Avenue

New York, NY 10021 (USA)

Phone Number:212-327-7987

Fax:212-327-7996

E-mail:gordon@morgan.rockefeller.edu, heath@fisher.rockefeller.edu,

ott@rockefeller.edu

Key words: Single Nucleotide Polymorphisms, Pedigree Errors, Probability Theory, Haplotype Relative Risk Sampling

Running Head: SNP True and Apparent Error Rate

Address correspondence to:

Derek Gordon

Laboratory of Statistical Genetics

Rockefeller University, Box 192

1230 York Avenue

New York, NY 10021 (USA)

Abstract. Single Nucleotide Polymorphisms (SNPs) are currently being developed for use in disequilibrium analyses. These SNPs consist of two alleles with varying degrees of polymorphism. A natural design for use with SNPs is the "haplotype relative risk" sampling design in which a father, mother, and child are typed at a SNP locus. Given such a trio of genotypes, we ask: what is the probability that a pedigree error (a change from one allele to the other) at a SNP locus will be detected using only Mendel's laws as a check? We calculate the probability of detecting such errors for a hypothetical SNP locus with varying degrees of polymorphism and for various true error rates. For the sets of allele frequencies considered, we find that the detection rates range between 25% - 30%, the detection rate being lowest when the two alleles have equal frequencies and the highest when one allele has a frequency of 10 %. Based on this detection rate, we determine that the true error rate is roughly 3.3 – 4 times that of the apparent error rate at a SNP locus. The greatest discrepancy between true and apparent error rates occurs when allele frequencies are equal.

1. Introduction

For disease genes of individually small effects, instead of genetic linkage analysis, disequilibrium analysis may be more promising [1]. Sampling designs have been developed ("haplotype relative risk" designs) in which the sampling unit is an affected offspring and his or her parents. Marker alleles transmitted to the offspring are contrasted to those not transmitted, where the latter form a contrived control sample [2, 3, 4]. Various efficient refinements of this approach have been developed [5, 6]. The data of interest for each of these tests are genotypes at a marker for a family trio, namely a father, mother, and a child.

One of the aims in developing marker genes of a new type (SNPs, Single Nucleotide Polymorphisms) is the detection of disease loci by disequilibrium analysis [7]. These SNPs have two alleles at each locus.

We consider the following problem: suppose that, for a random sample of families in which father, mother and child have been typed at a SNP locus, pedigree errors (any changes from one allele to another, e.g. due to sample swaps, genotyping error, etc.) occur randomly and independently with some fixed probability α. We say that an error in a trio is *undetectable* if the genotypes of the trio still display Mendelian consistency. The question we address is: what is the probability that errors in our sample of family trios go undetected? Let us label this quantity by β. We investigate the question for a single SNP locus with varying degrees of polymorphism. Once we calculate β, we can also calculate the detection rate, 1- β, and the apparent error rate, $\alpha(1-\beta)$. We illustrate the meaning of the expression apparent error rate with an example. Suppose we have 100 trios, so that we have 600 genotyped alleles. If the true error rate is $\alpha = 0.05$, or 5%, then 30 alleles are expected to have changed. If, say, the probability that an error goes undetected for $\alpha =$ 0.05 is $\beta = .40$, or 40%, the detection rate is $1 - \beta = 1 - 0.4 = 0.6$, or 60%. In this example, we will detect only 60% of the 30 errors, or 18 errors. Therefore, it will appear to us that the error rate in our set of genotypes is 18/600 = 0.03, or 3%. This value of 3% is what we call the apparent error rate for this example.

2. Methods

For a diallelic marker with allele numbers 1 and 2, there are three possible genotypes: 1/1, 1/2, and 2/2. Assuming Mendel's laws, there are exactly 15 ways in which parental

genotypes at a diallelic marker may be transmitted to a child. We list these possibilities in the Table 1. We define the ordered 3-tuple of genotypes

(Paternal Allele 1/Paternal Allele 2, Maternal Allele1/Maternal Allele 2, Child Allele 1/Child Allele 2)

in which the set of alleles is consistent with Mendel's laws, to be a *genotype trio*. In this definition we make no distinction between genotypes 2/1 and 1/2. Thus, we consider the trio (2/1, 2/1, 2/1) to be equal to (1/2, 1/2, 1/2), and so forth. For consistency, we shall always write the genotype 1/2. To distinguish between those 3-tuples that do and do not display Mendelian consistency, we shall use the term *general trio* to refer to *any* ordered 3-tuple of numbers in which each entry is either a 1 or a 2 allele, irrespective of whether the trio displays Mendelian consistency. From these definitions, we see that the set of genotype trios is a subset of the set of general trios. Also, we define the *conjugate* of a genotype trio M (denoted $\underline{M}$) to be the genotype trio that results when we replace each value of 1 in the trio M by a 2, and we replace each value of 2 in M by a 1. For example, the conjugate of the trio (1/1, 1/1, 1/1) is (2/2, 2/2, 2/2), the conjugate of (1/2, 1/1, 1/1) is (2/1, 2/2, 2/2), which we write as (1/2, 2/2, 2/2), and so forth. The list of all conjugates of genotype trios may be found in Table 1. We define an *error* in a genotype trio to be a change in one value of the trio. For example, if our original trio is (2/2, 2/2, 2/2) then we say we have introduced one error into that trio if we replace one and only one of the 2 alleles by a 1 allele. We extend our definition, in the obvious way, to speak of two or more errors in a genotype trio.

For each of the fifteen trios, we can introduce anywhere from 0 to 6 errors. In the case of either 0 or 6 errors, the resulting general trio will always display Mendelian consistency. We are interested in calculating, for a collection of genotype trios in which at

least one error has been introduced, the expected proportion of errors that go undetected. As above, we denote this quantity by β. Using basic probability theory [8], we have

$$\beta = \sum_{i=1}^{6} \Pr(\textit{undetected errors} \mid i \textit{ errors in trio}) \Pr(i \textit{ errors in trio}) \qquad (1)$$

We do not include $i = 0$ in our sum because we are only interested in the genotype trios in which errors have been introduced (β is a conditional probability). If we define $B(\alpha; i)$ by

$$B(\alpha; i) = \binom{6}{i} \alpha^i (1-\alpha)^{6-i},$$

then, because we assume that error introduction is random and independent for each allele in a genotype trio, and because we are only considering those genotype trios that contain at least one error, the quantity Pr(*i errors in trio*) in formula (1) is given by

$$\Pr(i \textit{ errors in trio}) = \frac{\binom{6}{i} \alpha^i (1-\alpha)^{6-i}}{1-(1-\alpha)^6} = \frac{B(\alpha; i)}{\sum_{i=1}^{6} B(\alpha; i)}.$$

Note that the expression $B(\alpha;i)$ is the probability density function, evalulated at i, $1 \le i \le 6$, for a binomial distribution with constant success rate α in each of 6 independent experiments.

Next, we calculate the quantity Pr(*undetected errors* | *i errors in trio*) in formula (1) . Using basic probability theory, we have

$$\Pr(\textit{undetected errors} \mid i \textit{ errors in trio}) = \sum_{M \in S} \Pr(N_0 \mid M, i) \Pr(M), \qquad (2)$$

where S is the set of all fifteen genotype trios (listed in the first column of Table 1), N_0 is the event that all errors in a trio go undetected, $\Pr(N_0 \mid M,i)$ is the probability that i errors

introduced into a genotype trio M go undetected, and Pr(M) is the probability of occurrence of the genotype trio M in a population of genotype trios. Pr (M) is calculated using the allele frequencies p for allele 1 and q (= $1 - p$) for allele 2. For each genotype trio M in Table 1, we calculate Pr(M) in Table 3. For each M and each i, the probability Pr($N_0 \mid M,i$) is equal to the proportion of resulting trios that, after i errors have been introduced, are consistent with Mendel's laws. For example, if M = (1/1, 1/1, 1/1) and i = 1, Pr($N_0 \mid M,i$) = 4/6, or 2/3; of the six trios that result from a change in one of the alleles, a change in any of the four parental alleles will display Mendelian consistency and a change in either of the two child alleles will display Mendelian inconsistency.

We now state some lemmas whose application simplifies the calculation of Pr($N_0 \mid M,i$) for any of the fifteen genotype trios M in Table 1 and any i, $0 \le i \le 6$. For the interested reader, we present proofs in the appendix. In each of these lemmas, N_0 is the event that no errors are detected in a genotype trio.

Lemma 1. For any genotype trio M and for any i, $0 \le i \le 6$, Pr($N_0 \mid M,i$) = Pr($N_0 \mid \underline{M},i$).

Lemma 2. For any genotype trio M and for any i, $0 \le i \le 6$,

Pr($N_0 \mid M,i$) = Pr($N_0 \mid M,6 - i$).

Lemma 3. For genotype trios M = (a/b, c/d, e/f) and N = (c/d, a/b, e/f),

Pr($N_0 \mid M,i$) = Pr($N_0 \mid N,i$) for any i , $0 \le i \le 6$.

In Table 2, we list the probability of detecting i errors, $1 \le i \le 3$, for each of six genotype trios. We determine these probabilities for each genotype trio M by listing all possible general trios that result when i errors have been introduced into M, and counting the subset of general trios that show Mendelian consistency. Probabilities for the remaining

trios in Table 1 can be calculated using Lemmas 1 and 3. For $4 \le i \le 5$, we can determine $\Pr(N_0 \mid M,i)$ using Lemma 2.

Given the probabilities in Table 2 and Lemmas 1 – 3, we are now able to calculate the value $\Pr(\textit{undetected error} \mid i \textit{ errors})$ in formula (2). For each value of i, this probability will be the sum of the entries under the appropriate heading in the column $\Pr(N_0 \mid M,i)\Pr(M)$ in Table 3. We see that:

$$\begin{aligned}
\textstyle\sum \Pr(N_0 \mid M,1)\Pr(M) &= 2/3\, p^4 + 3\, p^3q + 14/3\, p^2q^2 + 3\, pq^3 + 2/3\, q^4, \\
\textstyle\sum \Pr(N_0 \mid M,2)\Pr(M) &= 4/5\, p^4 + 14/5\, p^3q + 4\, p^2q^2 + 14/5\, pq^3 + 4/5\, q^4, \\
\textstyle\sum \Pr(N_0 \mid M,3)\Pr(M) &= 3/5\, p^4 + 14/5\, p^3q + 22/5\, p^2q^2 + 14/5\, pq^3 + 3/5\, q^4.
\end{aligned} \tag{2.2a}$$

Applying Lemma 1, we observe the remarkable fact that

$$\begin{aligned}
\textstyle\sum \Pr(N_0 \mid M,4)\Pr(M) &= \textstyle\sum \Pr(N_0 \mid M,2)\Pr(M), \\
\textstyle\sum \Pr(N_0 \mid M,5)\Pr(M) &= \textstyle\sum \Pr(N_0 \mid M,1)\Pr(M).
\end{aligned} \tag{2.2b}$$

We substitute the equations (2.2a) and (2.2b) into formula (2), which we then substitute into formula (1) to determine our desired quantity β, the probability that errors in a randomly sampled genotype trio go undetected. In Table 4, we calculate this quantity for various values of α and p.

3. Results and Discussion

From Table 4, we observe that, if we use only Mendel's laws to check whether or not a pedigree error has occurred, then the true error rate is roughly 3.3 to 4 times the magnitude of the apparent error rate. While this magnitude might not appear to be much for very small values of α, the effect becomes quite significant when one looks at apparent error rates of .05 or more. From results in Table 4, an apparent error rate of approximately .06 corresponds to a true error rate of approximately .20, regardless of the

allele frequencies used. Note that for each value of the true error rate α, the apparent error rates for both sets of allele frequencies are equal to two decimal places. Also note that, when one allele has frequency 0.1, the detection rate *decreases* from .3032 when α is .001 to .2865 when α is .300, a difference in detection rates of .0167. By contrast, when both allele frequencies are .5, the detection rate *increases* from .2501 when α is .001 to .2820 when α is .300, a difference in detection rates of .0319, almost twice that of the former difference.

As a comparison to the SNPs scenario, we performed simulations in which we generated 100,000 genotype trios for a marker with ten equally frequent alleles. We generated the simulated data using the SIMULATE program [9]. We created errors in the data by increasing a 1 allele to 2, decreasing a 10 allele to 9, and for every other allele, either increasing the allele by one or decreasing the allele by one, each with 50 % probability. As in the SNPs case, we assumed that the errors occurred independently and randomly at some rate α. For each setting of α, we generated a new set of 100,000 genotype trios. The results of these simulations are found in Table 5. We also calculate an empirical β by considering the quotient (total number of undetected errors)/(total number of errors introduced), and we report the empirical detection rate, $1 - \beta$.

Note that in the case of 10 equally frequent alleles, the lowest detection rate is .5850, when the true error rate is .0010. By contrast, for the same true error rate, in Table 4, using two equally frequent alleles, the detection rate is .2500. Comparing Tables 4 and 5, we see that in the SNPs case, the true error rate is approximately 3.5 - 4 times that of the apparent error rate when equal allele frequencies are used. In the case of 10 equally

frequent alleles, however, the true error rate is approximately 1.3 - 1.7 times the apparent error rate.

What are possible solutions to the problem of low detection rates for SNPs? One possible solution is to check for pedigree errors by performing multipoint analysis, and looking for multiple recombinations in an interval of less than 30cM [10]. However, with only two alleles at each of the SNP markers, the likelihood of observing true recombinations over a short interval is most probably reduced. The extent of this reduction needs to be researched.

Will genotyping additional siblings help? Possibly, but note that if genotyping is performed at only one locus, additional sibs can aid in finding errors in the parents, not in the first sibling. Also, if one of the alleles has a very high frequency, then the additional sib will provide even less information in determining errors. For example, consider the case where the 1 allele has a population frequency of 0.9. The majority of genotype trios will be (1/1, 1/1, 1/1). Let us assume that only one error occurs in this trio, as is the case when the true error rate is low. If it occurs in one of the parents, then the error goes undetected regardless of which sibling is typed. In the case where multiple loci are genotyped, additional sibs may determine phase in the parents and hence double recombinants in the first child. However, the extent to which genotyping additional sibs helps in finding SNPs errors also needs to be researched.

Given these results, we therefore suggest that researchers who look for pedigree errors in SNPs use additional methods beyond checking for Mendelian inconsistency. Suggestions include genotyping trios twice and genotyping multiple loci and looking for double recombinants.

4. Acknowledgments

This material is based upon work supported by the US Army Medical Research Acquisition Activity under award # DAMD17-94-J-4406. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the US Army Medical Research Acquisition Activity.

5. Appendix

In this section we present proofs of the three lemmas we stated in the body of the manuscript. We comment that in some of the proofs, we present equations that are equalities of *sets*. Briefly, two sets A and B are equal if they contain precisely the same elements. We shall use the notation $A \equiv B$ to indicate equality of sets A and B. Also, we shall use the notation $|A|$ to denote the cardinality, or size of the set A. For example, if the set A has three elements, then $|A| = 3$. Finally, if C represents a set of general trios, then we denote by $\underline{C}$ the set of all conjugates of elements of C. For further information on set theory notation and definitions, please see [11].

Lemma 1. For any genotype trio M and for any i, $0 \leq i \leq 6$, $\Pr(N_0 | M,i) = \Pr(N_0 | \underline{M},i)$.

Proof. Let $S(M, i)$ be the set of all general trios that result by introducing i errors into a genotype trio M. Also, let $U(M, i)$ be the subset of $S(M, i)$ of all general trios that are genotype trios, i.e., the trios that display Mendelian consistency. Using the identity $\underline{S(M, i)} \equiv S(\underline{M}, i)$, which is straightforward to check, and the fact that a general trio displays Mendelian consistency if and only if its conjugate displays Mendelian consistency, we have the identity $|U(M, i)| = |U(\underline{M}, i)|$. Using this identity and the fact that $|S(M, i)| = |S(\underline{M}, i)|$ we can write

$$\Pr(N_0 \mid M,i) = |U(M, i)|/ |S(M, i)| = |U(\underline{M}, i)|/ |S(\underline{M}, i)| = \Pr(N_0 \mid \underline{M},i),$$

which proves the lemma.

Lemma 2. For any genotype trio M and for any i, $0 \le i \le 6$, $\Pr(N_0 \mid M,i) = \Pr(N_0 \mid M,6 - i)$.

Proof. We prove this lemma by first proving the identity that, for any genotype trio M and any i, $0 \le i \le 6$,

$$\Pr(N_0 \mid M,i) = \Pr(N_0 \mid \underline{M},6 - i). \qquad (3)$$

To prove this identity, note first that we have the equality $S(M, i) \equiv S(\underline{M}, 6 - i)$. From this equality it follows immediately that $U(M, i) \equiv U(\underline{M}, 6 - i)$. If two sets are equal, then clearly they have the same cardinality. Thus, $|S(M, i)| = |S(\underline{M}, 6 - i)|$, $|U(M, i)| = |U(\underline{M}, 6 - i)|$, and

$$\Pr(N_0 \mid M,i) = |U(M, i)|/ |S(M, i)| = |U(\underline{M}, 6 - i)|/ |S(\underline{M}, 6 - i)| = \Pr(N_0 \mid \underline{M},6 - i),$$

so the identity is proved. Using identity (3), Lemma 1, and the fact that the conjugate of $\underline{M}$ is M, we have

$$\Pr(N_0 \mid M,i) = \Pr(N_0 \mid \underline{M},6 - i) = \Pr(N_0 \mid M,6 - i),$$

which proves Lemma 2.

Lemma 3. For genotype trios M = (a/b, c/d, e/f) and N = (c/d, a/b, e/f), $\Pr(N_0 \mid M,i) = \Pr(N_0 \mid N,i)$ for any i, $0 \le i \le 6$.

Proof. We have $|S(M, i)| = |S(N, i)|$. From the fact that a general trio M displays Mendelian consistency if and only if the general trio N, determined by reassigning the labels of father and mother in M, displays Mendelian consistency, it follows that $|U(M, i)| = |U(N, i)|$. From these two equalities and the fact that $\Pr(N_0 | A, i) = |U(A, i)| / |S(A, i)|$ for any genotype trio A and any i, $0 \leq i \leq 6$, the lemma follows.

Table 1. List of all genotype trios and their conjugates

Genotype Trio = M	Conjugate Genotype Trio $\underline{M}$
(1/1, 1/1, 1/1)	(2/2, 2/2, 2/2)
(1/1, 1/2, 1/1)	(2/2, 1/2, 2/2)
(1/1, 1/2, 1/2)	(2/2, 1/2, 1/2)
(1/2, 1/1, 1/1)	(1/2, 2/2, 2/2)
(1/2, 1/1, 1/2)	(1/2, 2/2, 1/2)
(1/1, 2/2, 1/2)	(2/2, 1/1, 1/2)
(1/2, 1/2, 1/1)	(1/2, 1/2, 2/2)
(1/2, 1/2, 1/2)	(1/2, 1/2, 1/2)
(1/2, 1/2, 2/2)	(1/2, 1/2, 1/1)
(2/2, 1/1, 1/2)	(1/1, 2/2, 1/2)
(1/2, 2/2, 1/2)	(1/2, 1/1, 1/2)
(1/2, 2/2, 2/2)	(1/2, 1/1, 1/1)
(2/2, 1/2, 1/2)	(1/1, 1/2, 1/2)
(2/2, 1/2, 2/2)	(1/1, 1/2, 1/1)
(2/2, 2/2, 2/2)	(1/1, 1/1, 1/1)

Table 2. Conditional probability that a genotype trio M displays Mendelian consistency if i errors are introduced, $1 \leq i \leq 3$

Trio = M	Pr(N_0 |M,1)	Pr(N_0 |M,2)	Pr(N_0 |M,3)
(1/1, 1/1, 1/1)	2/3	4/5	3/5
(1/2, 1/1, 1/1)	5/6	3/5	4/5
(1/2, 1/1, 1/2)	2/3	4/5	3/5
(1/1, 2/2, 1/2)	2/3	3/5	9/10
(1/2, 1/2, 1/1)	2/3	4/5	3/5
(1/2, 1/2, 1/2)	1	3/5	7/10

Table 3. Expected Proportion of Undetected Errors, Given that i errors are introduced into Genotype Trio, $1 \le i \le 3$

Trio = M	Pr(M)	Pr(N_0 |M,i)			Pr(N_0 |M,i) Pr(M)		
		i=1	i=2	i= 3	i=1	i=2	i=3
(1/1, 1/1, 1/1)	p^4	2/3	4/5	3/5	$(2/3)\,p^4$	$(4/5)\,p^4$	$(3/5)\,p^4$
(1/1, 1/2, 1/1)	p^3q	5/6	3/5	4/5	$(5/6)\,p^3q$	$(3/5)\,p^3q$	$(4/5)\,p^3q$
(1/1, 1/2, 1/2)	p^3q	2/3	4/5	3/5	$(2/3)\,p^3q$	$(4/5)\,p^3q$	$(3/5)\,p^3q$
(1/2, 1/1, 1/1)	p^3q	5/6	3/5	4/5	$(5/6)\,p^3q$	$(3/5)\,p^3q$	$(4/5)\,p^3q$
(1/2, 1/1, 1/2)	p^3q	2/3	4/5	3/5	$(2/3)\,p^3q$	$(4/5)\,p^3q$	$(3/5)\,p^3q$
(1/1, 2/2, 1/2)	p^2q^2	2/3	3/5	9/10	$(2/3)\,p^2q^2$	$(3/5)p^2q^2$	$(9/10)p^2q^2$
(1/2, 1/2, 1/1)	p^2q^2	2/3	4/5	3/5	$(2/3)\,p^2q^2$	$(4/5)p^2q^2$	$(3/5)\,p^2q^2$
(1/2, 1/2, 1/2)	$2\,p^2q^2$	1	3/5	7/10	$2\,p^2q^2$	$(6/5)p^2q^2$	$(7/5)\,p^2q^2$
(1/2, 1/2, 2/2)	p^2q^2	2/3	4/5	3/5	$(2/3)\,p^2q^2$	$(4/5)\,p^2q^2$	$(3/5)\,p^2q^2$
(2/2, 1/1, 1/2)	p^2q^2	2/3	3/5	9/10	$(2/3)\,p^2q^2$	$(3/5)p^2q^2$	$(9/10)p^2q^2$
(1/2, 2/2, 1/2)	pq^3	2/3	4/5	3/5	$(2/3)\,pq^3$	$(4/5)\,pq^3$	$(3/5)\,pq^3$
(1/2, 2/2, 2/2)	pq^3	5/6	3/5	4/5	$(5/6)\,pq^3$	$(3/5)\,pq^3$	$(4/5)\,pq^3$
(2/2, 1/2, 1/2)	pq^3	2/3	4/5	3/5	$(2/3)\,pq^3$	$(4/5)\,pq^3$	$(3/5)\,pq^3$
(2/2, 1/2, 2/2)	pq^3	5/6	3/5	4/5	$(5/6)\,pq^3$	$(3/5)\,pq^3$	$(4/5)\,pq^3$
(2/2, 2/2, 2/2)	q^4	2/3	4/5	3/5	$(2/3)\,q^4$	$(4/5)\,q^4$	$(3/5)\,q^4$
Total	1	Not Applicable			See Formulas (2.2a) and (2.2b)		

Table 4. Detection Rate 1 - β and apparent error rate $\alpha(1-\beta)$ for various values of true error rate α and various allele frequencies p at diallelic locus.

	Frequency of *1* allele = 0.1		Frequency of *1* allele = 0.5	
True error rate, α	Detection rate,1 -β	Apparent error rate, $\alpha(1-\beta)$	Detection rate, 1 -β	Apparent error rate, $\alpha(1-\beta)$
.0010	.3032	.0003	.2501	.0003
.0050	.3025	.0015	.2506	.0013
.0100	.3017	.0030	.2512	.0025
.0200	.3001	.0060	.2525	.0050
.0500	.2960	.0148	.2562	.0128
.1000	.2909	.0291	.2622	.0262
.2000	.2862	.0572	.2732	.0546
.3000	.2865	.0860	.2820	.0846

Table 5. Detection Rate 1 - β and apparent error rate $\alpha(1 - \beta)$ for various values of true error rate α at locus with 10 equally frequent alleles – Simulation Results.

True Error Rate, α	Empirical Detection Rate,1 - β	Apparent Error Rate, $\alpha(1 - \beta)$
.0010	.5850	.0006
.0050	.5900	.0030
.0100	.6010	.0060
.0200	.6050	.0121
.0500	.6290	.0310
.1000	.6630	.0660
.2000	.7160	.1430
.3000	.7510	.2250

References

1. Risch N, Merikangas K: The future of genetic studies of complex human diseases. Science 1996;273:516-1517.
2. Falk CT, Rubinstein P: Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 1987;51:227-233.
3. Ott J: Statistical properties of the haplotype relative risk. Genet Epidemiol 1989;6:127-130.
4. Thomson G: Mapping disease genes: family-based association studies. Am J Hum Genet 1995;57:487-498.
5. Terwilliger JD, Ott J: A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered 1992;42:337-346.
6. Spielman RS, Ewens WJ: The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 1996;59:983-989.
7. Collins FS, Guyer MS, Charkravarti A: Variations on a theme: cataloging human DNA sequence variation. Science 1997;278:1580-1581.
8. Hogg RV, Craig AT: Introduction to Mathematical Statistics 1978. Macmillan Publishing Company, New York.
9. Terwilliger JD, Ott J: Handbook of Human Genetic Linkage 1994:Johns Hopkins University Press, Baltimore.
10. Ott J: Analysis of Human Genetic Linkage 1991:Johns Hopkins University Press, Baltimore.
11. Hungerford TW: Algebra 1984:Springer-Verlag, New York.

A Principal Components Approach Based on Heritability for Combining Phenotype Information

Jurg Ott[1] and Daniel Rabinowitz[2]

[1]Laboratory of Statistical Genetics, Rockefeller University, New York and

[2]Department of Statistics, Columbia University, New York

Running title:

Principal Components of Heritability

Corresponding author:

Daniel Rabinowitz

Department of Statistics

Mathematics Building

Columbia University

New York, NY 10027

(212) 854-3400

fax (212) 663-2454

Summary

For many traits, genetically relevant disease definition is unclear. For this reason, researchers applying linkage analysis often obtain information on a variety of items. With a large number of items, however, the test statistic from a multivariate analysis may require a prohibitively expensive correction for the multiple comparisons. The researcher is faced, therefore, with the issue of choosing which variables or combinations of variables to use in the linkage analysis. One approach to combining items is to first subject the data to a principle components analysis, and then performs the linkage analysis the first few principle components. However, principal components analyses do not take family structure into account. Here, an approach is developed in which family structure is taken into account when combining the data.

The essence of the approach is to define principal components of heritability as the scores with maximum heritability in the data set, subject to being uncorrelated with each other. The principal components of heritability may be calculated as the solutions to a generalized eigensystem problem. Four simulation experiments are used to compare the power of linkage analyses based on the principal components of heritability and the usual principal components. The first of the experiments corresponds to the null hypothesis of no linkage. The second corresponds to a setting where the two kinds of principal components coincide. The third corresponds to a setting in which they are quite different and where the first of the usual principal components is not expected to have any power beyond the type I error rate. The fourth set of experiments corresponds to a setting where the

usual principal components and the principal components of heritability differ, but where the first of the usual principal components is not without power. The results of the simulation experiments indicate that the principal components of heritability can be substantially different from the standard principal components and that when they are different, substantial gains in power can result by using the principal components of heritability in place of the standard principal components in linkage analyses.

Introduction

For many traits, while there may exist unequivocal diagnostic schemes for particular disease entities, genetically relevant disease definition is unclear. An underlying gene or set of genes may, for example, be related to a very specific trait such a schizophrenia or lead to susceptibility to a spectrum of diseases such as schizophrenia, schizotypical disorder, and bipolar disease. For this reason, researchers applying linkage analysis with complex traits often obtain from each study subject information on a variety of items relevant to the trait in question. The information on these items may be coded as quantitative variables or as scales that combines information from several items. For example, Basset et al (1993) discuss the use of The Positive and Negative Syndrome Scale (PANSS) of Kay et al (1987) for characterizing schizophrenia in pedigree members.

Any of the variables or scales may be subjected to the desired genetic analysis and the results can then be compared heuristically. With a large number of items, however, such an exploratory approach may be unwieldy. Furthermore, the corrections needed to account for the multiple looks corresponding to the many items may be prohibitive. Alternatively, the items may used together as a multivariate outcome in a single analysis. See, for example, Schork (1993) and Amos et al. (1990). With a large number of items, however, the power of a multivariate analysis to detect linkage can be substantially lower than the power of an analysis applied to a genetically relevant scale. This is because the increased degrees of freedom for the reference distribution of a multivariate analysis can result in too stringent a

criterion for statistical significance.

One approach to choosing a scale is to subject the available items to a principal components analysis. The first few principle components are then candidates for linkage analyses. Lindström and von Knorring (1993) for example, subject PANSS scores to a principle components analysis. See also Lindenmayer et al (1995). An intermediate step may be to estimate the heritability of the first several principal components, and to use the components with highest heritability. See Hasstedt et al (1994) for an example concerning sodium transport and Livshits (1995) for an example concerning body size and shape. Farmer et al (1987) examined heritability for several different diagnostic criteria for schizophrenia. However, principle components depend on the variance covariance matrix of the data pooled from all pedigrees, and thus do not reflect family structure. By focusing only on scales obtained as principal components, other scales with higher heritability may be overlooked.

Here, an approach is developed in which the available items are combined into scales, not on the basis of their variance-covariance structure, but instead, on the basis of heritability. The first scale has highest heritability, the second scale has highest heritability among all scales uncorrelated with the first, the third has highest heritability among those uncorrelated with the first and second, and so on. Heritability is the ratio of the variances of family specific components and individual specific components of variance. Thus, a combination of the variance-covariance structure of both the between family and within family variance components are used to compute the scales.

Approaches similar in spirit have been contemplated previously. Zlotnik et al (1983) describe an approach based on choosing linear combinations to maximize the likelihood under the hypothesis of single gene segregation in a single pedigree. Principal components analysis was used in the computations. Multivariate selection indices are computed in the context of plant and animal breeding. Selection indices combine the economic value or relative fitness of traits together with heritability, and the computation of selection indices has parallels with the computations suggested in the next section. See, for example, Humphreys (1995), Hazel (1943) or Kempthorne (1969). Boomsma used simulation studies to examine the increase in power that can occur when using using linear combinations based on factor scores to detect linkage in the presence of pleitropy. Comuzzie et al. (1997) describe two similar approaches. The first involves "conditioning each phenotype on the common or shared genetic effects with the others in the group to maximize the variance of each trait attributable to unique loci". The goal of such an approach is to partial out the effect of different loci, while the goal of the approach presented here is to combine information in order to have greater power in the presence of pleitropy. The second of Comuzzie et al's (1997) approaches involves computing estimates of the principal components of the genetic portion of the phenotype data. The the approach presented here differs in that it involves computing what might be though of as the principal components of the genetic portion of the data relative to the residual variability in the phenotypes. In the next section, the approach to choosing scales is detailed. Also described in the next section are simulation

experiments that investigate the relative efficiency of linkage analyses based on the approach. In the third section, the results of the simulation experiments are reported.

Methods

It is convenient to focus on the situation in which phenotype data in the form of a number of items have been obtained from the offspring in a sample of nuclear families. Let p denote the number of items, and let Y denote the p dimensional vector of items. It is natural to think of the vector of items as composed of a family specific component, A, and a individual specific component, E,

$$Y = A + E,$$

with the two components uncorrelated with each other. The family specific component is induced by variation in the parental genes, and by shared environmental factors. The individual specific component is induced by subject specific environmental factors, and differences in transmission of parental alleles. Let Σ_A and Σ_E denote the variance-covariance matrices of A and of E, respectively. Then the heritability of a linear combination of the the items $c^T Y$ may be expressed as

$$c^T \Sigma_A c \Big/ c^T \left(\Sigma_A + \Sigma_E\right) c \,,$$

where the column vector c is the coefficients of the linear combination, and c^T is its transpose.

The standard principal components are defined as the scores with maximum variance, subject to being uncorrelated with each other. See, for example, Rao (1964) or Morrison (1976). The principal components of heritability are defined not as the scores with maximum variance, but instead, as the scores with maximum heritability, subject to being

uncorrelated with each other. That is, the principal components of heritability are scores $c_1^T Y$, $c_2^T Y$, ..., $c_p^T Y$ with the property that: $c_1^T Y$ has maximum heritability; $c_2^T Y$ has maximum heritability subject to $c_2^T Y$ being uncorrelated $c_1^T Y$; $c_3^T Y$ has maximum heritability subject to $c_3^T Y$ being uncorrelated with $c_1^T Y$ and $c_2^T Y$; and so on. It may be of interest to note that not only are the principal components of heritability uncorrelated, but that their family specific components, the $c_i^T A$, are uncorrelated as well.

It is well known that c_1, c_2, ... and c_p are the solutions to the generalized eigensystem problem

$$\Sigma_A c = \lambda \Sigma_E c.$$

Press et al (1988) discuss numerical methods for solving the generalized eigensystem problem based on an approach described by Wilkinson and Reinsch (1971). Given estimates of the variance-covariance matrices, the principal components of heritability may be estimated by solving the generalized eigensystem problem for the estimated matrices.

In order to examine the relative efficiency of using the principal components of heritability in linkage analyses, four simulation experiments were carried out. In all of the experiments, there were five items from each of two children in nuclear families. The genetic model had a single locus influencing the trait. The locus had two equally prevalent alleles. The effect of the alleles on the traits was additive: presence of a copy of the first of the two alleles added a constant to each of the items; the other allele had no effect. The constant was the same across families, but differed across items. The items were then generated by adding a multivariate normal vector to the

effect of the gene. Thus, the family specific component of variance was induced by the parental alleles at the gene, and the subject specific variance was induced by the multivariate normal residual variance together with variation in the siblings' genetic makeup around the conditional expectation given the parental alleles.

The procedure of Haseman and Elston (1972) was used to test for linkage. A completely informative marker perfectly linked to the underlying trait locus was simulated. In all of the experiments, the procedure was applied to both the estimated first standard principal component and the estimated first principal component of heritability. That is, the squared differences in the first of each kind of principal component were separately regressed on the number of alleles shared identical by descent by the two siblings. The regression coefficient, normalized by its standard error, was used as the test statistic to assess significance. In each of the experiments, 10,000 replications were performed.

In order to compute the principal components, estimates of the family specific and subject specific variance-covariance matrices are required as input to the generalized eigensystem problem. Let Y_{ij} denote the column vector of items from the j^{th} sibling in the i^{th} family. Then, the estimates of the variance covariance matrices of the subject specific component of variance and the family specific component of variance used in the simulations may be expressed as

$$\widehat{\Sigma}_E = \frac{1}{\sum_{i=1}^{n}(m_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^{\otimes 2},$$

and

$$\hat{\Sigma}_A = \frac{1}{\sum_{i=1}^{n} m_i - 1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left(Y_{ij} - \bar{Y}_{..}\right)^{\otimes 2} - \hat{\Sigma}_E$$

where $\bar{Y}_{..}$ is the average over all siblings in the study, and $\bar{Y}_{i.}$ is the average over the m_i siblings in the i^{th} family. The notation $V^{\otimes 2}$ denotes the matrix multiplication of the column vector V by its transpose, $V^{\otimes 2} = VV^T$. All of the calculations, including random number generation, were carried out in FORTRAN77 using the IMSL (1994) subroutines library.

The models for the simulation experiments are described in the first column of table 1. In the first experiment, there was no genetic effect at all, and the items were independent and identically distributed. Each replication had 250 families. In the second simulation, the vector added to the phenotype through presence of one of the alleles had all of its components equal to 1. The multivariate normal component was independent identically distributed variables with expectation equal to 0 and variance equal to 1. The number of families in the second experiment was 600. In the third experiment, the vector added had 1.5 in its first two components, and 0 in the other three components. The variance covariance matrix of the multivariate normal observation had 2.0 in the last three components of the diagonal, and 0.25 in the first two. The covariances between the last three items was 0.5, while the first two items were assumed uncorrelated with all the others. The number of families was 150. In the fourth experiment, the first and last components of the multivariate normal contribution to the items had covariance 1.0. All other covariances were zero. The last component had variance 1.5, while all the others had variance 1. The vector

added by presence of the allele was 1.0 in the first component, and zero in the others. The number of families was 80.

The first principal component and the first principal component of heritability for the models corresponding to the simulation experiments are listed in in the second and third columns of table 1. The matrices Σ_A and Σ_E are computed as $\mu\mu^T/4$ and $\mu\mu^T/4 + \Sigma$, where Σ is the variance covariance matrix of the multivariate normal component. From the heritabilities, it is expected that that linkage analyses based on the principal components of heritability should improve on analyses based on the usual principal components in the third and fourth experiments, while it is hoped that the power of both approaches will be comparable in the second experiment. Neither approach is expected to have power in the first experiment.

Results

The results of the simulations experiments are recorded in table 2. The rows of the table correspond to different significance levels, and the entries in the table are the empirical probability of detecting linkage. The columns of the table correspond to the results for the principal components of heritability and the standard principal components for the four scenarios described in the previous section.

The first set of experiments had no genetic effect. The power is as expected for both scales: it is the nominal α level.

The second set of experiments corresponds to a situation in which the first principal component and the first principal component of heritability are the same. The standard principal component analysis has somewhat greater power than the principal component of heritability analysis, however. An examination of the estimated coefficients of the principal components provides an explanation: the coefficients of the principal components of heritability are estimated with somewhat less accuracy than the principal components.

The third simulation experiment corresponds to a situation in which the principal components of heritability are very different from the standard principal components. In this setting, the variability due to genetic causes is in the first two items. The first of the standard principal components, however, is a linear combination of the last three components. It is not surprising that the analysis based on the principal components of heritability improves, in this setting, on the analysis based on the standard principal

components.

The fourth simulation experiment is intermediate between the second and the third. The first of the standard principal components puts weight on the first item, while the first of the principal components of heritability places weight on only the first item. As expected, the power of the analysis based on the principal component of heritability is greater than that of the analysis based on the standard principal component. However, the analysis based on the standard principal component is not without power.

Discussion

In many applications, there may be additional environmental or subject specific covariates that should be included in the analyses. When this is the case, it would be reasonable to first regress out the covariates before estimating the covariance structure. Perhaps preferably, a mixed effects regression model could be used to estimate the regression coefficients and the variance components simultaneously. See, for example, Laird and Ware (1983).

Here, attention has focussed on sib-pair analyses and additive genetic models. With general pedigree data, modeling and estimation of the family and individual specific components of variance are more complicated. However, in principle, the approach proposed here is applicable: after estimation of the components of variance, the principle components of heritability may be estimated by solving the generalized eigensystem problem. In any case, crude estimates of the variance covariance structures may be obtained by treating all individuals within a pedigree the same. A variety of more sophisticated techniques for computing variance components in pedigrees have been devised. See, for example, Blangero and Konigsberg (1991).

Issues related to ascertainment have not been discussed. For a monogenic group of traits, if pedigrees are ascertained through affected individuals, the family specific component of variance can be, in the data set, roughly constant. In such settings, especially if variability due to genetic variability within a pedigree are a large portion of the subject specific component of

variance, then the usual principle components may be preferable to the principle components of heritability.

With only pedigree structure and phenotype information, but without genotype information, it is difficult to disentangle the effects of shared environmental factors from the effect of shared genetic material, and it can be impossible to disentangle the effects of shared genetic material at different loci. Therefore, obtaining scales on the basis of heritability exposes one to the risk that the scales are not relevant to any particular locus, but rather to some combination of loci and environmental factors. In fact, it is possible that in some situations, environmental factors may induce family specific components of variance that lead to principal components of heritability that have no power for detecting linkage, while the standard principal components would have at least some power. See, for example Jiang and Zeng (1995) for a discussion of this issue. Nevertheless, it seems natural that, if scales are to be formed without reference to genotype information, then the scales are most likely best developed with reference to heritability. A possible strategy when faced with more than a few phenotype variables might be to first apply a linkage analysis to the first or the first few principal components of heritability. If linkage is not detected with the principal component, then a multivariate linkage analysis might be performed, even if the large degrees of freedom associated with a multivariate analysis with many phenotype variables would most likely preclude a statistically significant result.

Conclusions

The results of the simulation experiments indicate that in some situations, the principal components of heritability can be substantially different from the standard principal components. Furthermore, when the principal components of heritability are used in the place of the standard principal components, substantial gains in power can result. The third of the simulation experiments is an example in which the standard principal component would be a much poorer choice than the principal component of heritability.

However, the results of the simulation experiments also indicate that some care must be taken in settings where the estimate of the subject specific components of variance is unstable. In such cases, the instability in the estimate of the principal components of heritability may outweigh their potential greater power. This will especially be the case in settings where the principal components of heritability and the standard principal components are not very different.

Acknowledgments

This work was supported by grant GM55978 (to Daniel Rabinowitz) from the National Institutes of General Medical Sciences, and by the US Army Medical Research Acquisition Activity under award # DAMD17-94-J-4406 (to Jurg Ott). Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the US Army Medical Research Acquisition Activity. The authors thank Dr. Simon Heath for pointing out the relationship to selection indices.

References

Amos CI, Elston RC, Bonney GE, Keats BJB and Berenson GS (1990) A multivariate Method for Detecting Genetic Linkage, with Application to a Pedigree with an Adverse Lipoprotein Phenotype. American Journal of Human Genetics **47**247-254.

Basset AS, Collins EJ, Nuttall SE and Honer WG (1993) Positive and negative symptoms in families with schizophrenia. Schizophrenia Research **11**:9-19.

Bell MD, Lysaker PH, Beam-Goulet JL, Milstein RM and Lindenmayer JP (1994) Five-component model of schizophrenia: assessing the factorial invariance of the positive and negative syndrome scale. Psychiatry Research **52**:295-303.

Blangero J amd Konigsberg LW (1991) Multivariate segregation analysis using the mixed model. Genetic Epidemiology **8**: 299-316.

Boomsma DI (1996) Using multivariate genetic modeling to detect pleiotropic quantitative trait loci. Behav. Genet. **26**:161-166.

Comuzzie AG, Mahaney MC, Almasy L, Dyer TD and Blangero J (1997) Exploiting Pleiotropy to Map Genes for Oligogenic Phenotypes using Extended Pedigree Data. Genetic Epidemiology **14**:975-980.

Farmer AE, McGuffin P and Gottesman II (1987) Twin concordance for DSM-III schizophrenia. Scrutinizing the validity of the definition. Archives of General Psychiatry **44**:634-41.

Haseman JK and Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behavioral Genetics **2**: 3-19.

Hasstedt SJ, Hunt SC, Wu LL and Williams RR (1994) Evidence for multiple genes determining sodium transport. Genetic Epidemiology **11**:553-68.

Hazel LN (1943) The genetic basis for constructing selection indexes. *Genetics* **28**:476-490.

Humphreys MO (1995) Multitrait response to selection in *Lolium perenne* L. (perennial ryegrass) populations. Heredity **74**:510-517.

IMSL 1994 FORTRAN subroutines for statistical applications and FORTRAN subroutines for mathematical applications. Visual Numerics, Inc., Houston.

Jiang C and Zeng ZB (1995) Multiple Trait Analysis of Genetic Mappiung for Quantitative Trait Loci. Genetics **140**:1111-1127.

Kay SR and Sandyk R (1987) Experimental models of schizophrenia. International Journal of Neuroscience. **58**:69-82.

Kempthorne O (1969) An introduction to genetic statistics. Iowa State University Press:Ames Iowa.

Laird N and Ware J (1983) Random-Effects Models for Longitudinal Data. Biometrics **38**:963-974.

Lindenmayer JP, Bernstein-Hyman R, Grochowski S and Bark N. (1995) Psychopathology of Schizophrenia: initial validation of a 5-factor

model. Psychopathology **28**:22-31.

Lindström E and von Knorring L. (1994) Symptoms in schizophrenic syndromes in relation to age, sex, duration of illness and number of previous hospitalizations. Acta Psychiatrica Scandinavica **89**:274-8.

Livshits G, Otremski I and Kobyliansky E (1995) Genetics of human body size and shape: complex segregation analysis. Annals of Human Biology **22**:13-27.

Morrison DF (1976) Multivariate Statistical Methods, 2nd Edition. McGraw-Hill Book Co., New York.

Press WH, Fannery BP, Teukolsky SA and Vetterling WT (1988) Numerical Recipes in C. Cambridge University Press, Cambridge.

Rao, CR (1964) The use and Interpretation of Principal Component Analysis in Applied Research. Sankhya A **26**:329-358.

Schork NH (1993) Extended multipoint identity -by-descent analysis of human quantitative traits; efficiency, power and modelling considerations. American Journal of Human Genetics. **53**:1306-1319.

Wilkinson JH and Reinsch C (1971) Linear Algebra, vol. II of Handbook for Automatic Computation. Springer-Verlag, New York.

Zlotnik LH, Elston RC and Namboodiri KK (1983) Pedigree discriminant analysis: a method to identify monogenic segregation. American Journal of Medical Genetics **15**:307-13.

Table 1.

Models for the simulation experiments.

The observations for an individual are denoted by Y, the number of transmitted first alleles at the trait locus is denoted by X. The notation $N(0, \Sigma)$ refers to a multivariate normal vector with expectation 0 and variance-covariance matrix Σ. The coefficients of the principal components of heritability and principal components, together with their heritabilities are also recorded. The models are listed in the same order as described in the text.

Model	PCH, Heritability	PC, Heritability
$Y = N\left(0, \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}\right)$	Doesn't exist	$\begin{pmatrix} 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \end{pmatrix}, 0$
$Y = X\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + N\left(0, \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}\right)$	$\begin{pmatrix} 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \end{pmatrix}, 0.556$	$\begin{pmatrix} 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \end{pmatrix}, 0.556$
$Y = X\begin{pmatrix} 1.5 \\ 1.5 \\ 0 \\ 0 \\ 0 \end{pmatrix} + N\left(0, \begin{pmatrix} 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 2.0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 2.0 & 0.5 \\ 0 & 0 & 0.5 & 0.5 & 2.0 \end{pmatrix}\right)$	$\begin{pmatrix} 0.707 \\ 0.707 \\ 0 \\ 0 \\ 0 \end{pmatrix}, 0.667$	$\begin{pmatrix} 0 \\ 0 \\ 0.577 \\ 0.577 \\ 0.577 \end{pmatrix}, 0$
$Y = X\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + N\left(0, \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1.5 \end{pmatrix}\right)$	$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, 0.2$	$\begin{pmatrix} 0.309 \\ 0.541 \\ 0.541 \\ 0.541 \\ 0.190 \end{pmatrix}, 0.032$

Table 2.

Empirical probability, from simulation experiments with 10,000 replications, of detecting linkage when the Haseman-Elston procedure is applied to the first principal component (PC) and to the first principal component of heritability (PCH) in four different scenarios. The four scenarios are described at the end of the methods section.

	First		Second		Third		Fourth	
α level	PCH	PC	PCH	PC	PCH	PC	PCH	PC
0.100	0.103	0.105	0.706	0.806	0.998	0.144	0.890	0.543
0.050	0.051	0.051	0.577	0.694	0.996	0.085	0.784	0.349
0.010	0.011	0.010	0.303	0.405	0.969	0.031	0.438	0.082
0.005	0.005	0.006	0.211	0.301	0.928	0.023	0.291	0.039
0.001	0.002	0.001	0.082	0.128	0.744	0.011	0.094	0.006

Some key words: complex traits, factor analysis, genetics, multivariate, segregation analysis, variance components.

11 September 1998

User's Guide to the EH program

Jurg Ott, Rockefeller University, New York (ott@complex.rockefeller.edu)

Introduction

EH is a linkage utility program to test and estimate linkage disequilibrium between different markers or between a disease locus and markers.

1) In the first case (no disease), the data consist of the number of individuals for each genotype that are taken to be collected at random from a population. Based on these sample data, the EH program estimates allele frequencies for each marker. Haplotype frequencies are estimated with allelic association (H_1) and without (H_0). The EH program also provides log likelihood, chi-square and the number of degrees of freedom under hypotheses H_0 and H_1.

2) In the second case (with disease), there are two data sets, a sample of affected individuals (case sample) and a control sample. Each sample comprises unrelated individuals. Given disease gene frequency, penetrance of each genotype, and the case-control data, the EH program first estimates the allele frequencies for each marker assuming no allelic association. It then estimates the haplotype frequencies under the assumptions of no allelic association (H_0), allelic association among markers but not with the disease (H_1), and allelic associations among all loci (H_2). The EH program also calculates log likelihood, chi-square and the number of degrees of freedom under each of the hypotheses, H_0, H_1 and H_2. For more information please refer to Terwilliger and Ott (1994).

Current program version: Two separate program constants are distinguished, one for maximum number of haplotypes and one for maximum number of genotypes, ie. the product over the number of genotypes at each locus. In DOS, maxhap = 500 and maxgeno = 1800. In addition, as before, there are constants for maximum number of loci and maximum number of alleles at each locus. Also, the program now reports errors in words rather than just in the form of an error number. The EH program is available in a DOS and a Unix version.

Note: For sparse data (relatively few observations with large numbers of alleles), the chi-square approximation to the test statistics used by EH is unreliable. Then, more sophisticated programs such as Arlequin are appropriate (http://anthropologie.unige.ch/arlequin/).

Files in this package (DOS):

EH.TXT: This documentation.

EH.PAS: Source code of EH program.

EH.EXE: Executable code of EH program, which is compiled with a maximum of 30 alleles per locus, 10 loci, and 650 haplotypes.

CASE.DAT, CONTROL.DAT, EH.DAT: Sample input data files.

Installation

The EH program does not require additional programs although you need a Pascal compiler (Borland Pascal or Turbo Pascal) to recompile the program when you change program constants. The executable file may be put into any directory as long as that directory is accessible. EH was developed for Turbo/Borland Pascal and has recently been ported to Unix.

Depending on how it is used (mode 1 or 2, above), the EH program requires one or two input files (see Terwilliger and Ott, 1994). They may be named on input (interactively). Below, the files are described under their default names.

1) No disease: There is one input file, EH.DAT. It contains the number of alleles for each marker and the observations for each genotype.

First line:
Number of alleles at the first marker, number of alleles at the second marker, and so on. Assuming you have 2 markers, the first marker has 2 alleles and the second marker has 3, you write 2 3 in the first line. The order of markers in the remainder of the input file is determined by the order of markers you entered in the first line.

Subsequent lines:
Number of observations for given genotypes. These numbers must be arranged as follows:
The number of columns is the number of the possible genotypes at the last locus. Let M be the number of alleles at the last locus, then the number of the possible genotypes equals M(M+1)/2. For example, if the last locus has two alleles, then there are 3 possible genotypes which are 1/1, 1/2 and 2/2. Therefore, in each line there are 3 columns with the order of 1/1, 1/2, 2/2. Similarly, if the last marker has three alleles, then there are 6 columns in each line corresponding to 1/1, 1/2, 2/2, 1/3, 2/3, 3/3.
The number of rows is the product of the number of the possible genotypes at the first (N-1) markers, where N is the total number of markers. That is, no. of rows = $L_1(L_1+1)/2 \times L_2(L_2+1)/2 \times \ldots \times L_i(L_i+1)/2 \ldots$, where L_i is the number of alleles at the i-th locus.

For example, assume you have 3 loci and the first and the second locus each have 2 alleles, and the third locus has 3 alleles. Excluding the first row, there are 6 columns for each row. Also, there are 6 rows in addition to the first row. However, if the first locus has 3 alleles and the second and third have 2 alleles each, then in addition to the first row, there are 18 rows and 3 columns for each row. Refer to the examples section for details.

2) For a Case-Control study, there are two data files, CASE.DAT and CONTROL.DAT. CASE.DAT contains the data for the affected individuals and CONTROL.DAT contains the data for the unaffected individuals. The format of these two data files is the same as that for the random sample data. Information on disease inheritance (allele frequency, penetrances) must be furnished interactively. Based on the known disease gene frequency, penetrance of each genotype, and CASE.DAT and CONTROL.DAT, haplotype frequencies are estimated.

The output file from the EH program, EH.OUT, contains the estimated haplotype frequencies and their corresponding log likelihoods.

Running the EH program

To run the EH program simply type EH at the DOS prompt. Below, three examples are provided, two for random sample data and one for case-control data.

Example 1:

Assume we have a random sample with three markers, 2 alleles each for the first and the second markers, and 3 alleles for the third marker. The observations for each genotype are as follows:

		Locus 3					
Locus 1	Locus 2	1/1	1/2	2/2	1/3	2/3	3/3
1/1	1/1	0	0	0	12	3	9
1/1	1/2	2	12	2	2	3	4
1/1	2/2	5	4	4	2	0	18
1/2	1/1	7	2	0	2	6	3
1/2	1/2	9	0	0	3	4	2
1/2	2/2	10	3	0	2	3	8
2/2	1/1	1	2	10	2	9	4
2/2	1/2	5	3	4	6	7	3
2/2	2/2	9	3	0	0	3	10

According to the instructions in the input file section, we set up the input file as follows:

```
 2  2  3
 0  0  0 12  3  9
 2 12  2  2  3  4
 5  4  4  2  0 18
 7  2  0  2  6  3
 9  0  0  3  4  2
10  3  0  2  3  8
 1  2 10  2  9  4
 5  3  4  6  7  3
 9  3  0  0  3 10
```

The output file from the EH program is:

```
Estimates of Gene Frequencies (Assuming Independence)
----\-----------------------------------
locus \ allele     1       2       3
--------\---------------------------------
   1    |         0.5022  0.4978
   2    |         0.4736  0.5264
   3    |         0.3436  0.2357  0.4207
----------------------------------------
# of Typed Individuals: 227

There are 12 Possible Haplotypes of These 3 Loci.
They are Listed Below, with their Estimated Frequencies:
-------------------------------------------------------------
| Allele   Allele   Allele  |      Haplotype Frequency      |
|   at       at       at    |                               |
| Locus 1  Locus 2  Locus 3 |  Independent   w/Association  |
```

```
-------------------------------------------------------------
   1         1         1          0.081720     0.091796
   1         1         2          0.056052     0.020709
   1         1         3          0.100055     0.115110
   1         2         1          0.090843     0.078947
   1         2         2          0.062309     0.067771
   1         2         3          0.111224     0.127870
   2         1         1          0.081004     0.055169
   2         1         2          0.055560     0.117023
   2         1         3          0.099177     0.073761
   2         2         1          0.090046     0.117700
   2         2         2          0.061762     0.030180
   2         2         3          0.110248     0.103964
-------------------------------------------------------------
# of Iterations = 14

                                      #param   Ln(L)      Chi-square
-------------------------------------------------------------------
H0: No Association                        4   -953.90       0.00
H1: Allelic Associations Allowed         11   -934.98      37.84
```

Example 2:

Again we have three markers, but with different numbers of alleles. The first marker has 3 alleles, and the second and the third each have 2 alleles. The observations for each genotype are as follows:

```
-------------------------------
                      Locus 3
                     -----------
Locus 1   Locus 2    1/1 1/2 2/2
           1/1        0   2   5
   1/1     1/2        7   9  10
           2/2        1   5   9
           1/1        0  12   4
   1/2     1/2        2   0   3
           2/2        2   3   3
           1/1        0   2   4
   2/2     1/2        0   0   0
           2/2       10   4   0
           1/1       12   2   2
   1/3     1/2        2   3   2
           2/2        2   6   2
           1/1        3   3   0
   2/3     1/2        6   4   3
           2/2        9   7   3
           1/1        9   4  18
   3/3     1/2        3   2   8
           2/2        4   3  10
-------------------------------
```

Data input looks as follows:

```
3 2 2
   0   2   5
   7   9  10
   1   5   9
   0  12   4
   2   0   3
   2   3   3
   0   2   4
   0   0   0
```

```
 10   4   0
 12   2   2
  2   3   2
  2   6   0
  3   3   0
  6   4   3
  9   7   3
  9   4  18
  3   2   8
  4   3  10
```

The estimated allele frequencies and haplotypes are the same as in example 1.

Example 3:

Example for case-control study. The CASE data consist of the affected individuals and the CONTROL data comprise the unaffected individuals. The user must provide disease gene frequency and penetrance for each genotype. Assume a dominant disease with gene frequency 0.01 and penetrances 0, 1, 1 for the respective genotypes +/+, D/+, and D/D.

Case data set:

```
     1/1   1/2   2/2
1/1   10    20    10
1/2   20    40    20
2/2   10    20    10
```

Control data set:

```
      1/1  1/2  2/2
1/1    1    2    1
1/2    8   16    8
2/2   16   32   16
```

Output based on the above data sets and disease information is as follows:

```
Estimates of Gene Frequencies (Assuming Independence)
(Disease gene frequencies are user specified)
----\-------------------------
locus \ allele     1       2
--------\-----------------------
Disease |         0.9900  0.0100
   1    |         0.3846  0.6154
   2    |         0.5000  0.5000
---------------------------------
# of Typed Individuals: 260

There are 8 Possible Haplotypes of These 3 Loci.
They are Listed Below, with their Estimated Frequencies:
-------------------------------------------------------------------------
| Allele   Allele   Allele  |         Haplotype   Frequency          |
|   at       at       at    |                                        |
| Disease  Marker1  Marker2 |  Independent   Ind-Disease   w/Asso.   |
-------------------------------------------------------------------------
   +        1        1          0.190385      0.190385     0.125060
   +        1        2          0.190385      0.190385     0.125060
   +        2        1          0.304615      0.304615     0.369940
   +        2        2          0.304615      0.304615     0.369940
```

D	1	1	0.001923	0.001923	0.003397
D	1	2	0.001923	0.001923	0.003397
D	2	1	0.003077	0.003077	0.001603
D	2	2	0.003077	0.003077	0.001603

# of Iterations = 1

	#param	Ln(L)	Chi-square
H0: No Association	2	-539.16	0.00
H1: Markers Asso., Indep. of Disease	3	-539.16	0.00
H2: Markers and Disease Associated	6	-539.16	0.00

There are 8 Possible Haplotypes of These 3 Loci.
They are Listed Below, with their Estimated Frequencies:

Allele at Disease	Allele at Marker1	Allele at Marker2	Haplotype Frequency: Independent	Ind-Disease	w/Asso.
+	1	1	0.190385	0.190385	0.124413
+	1	2	0.190385	0.190385	0.124413
+	2	1	0.304615	0.304615	0.370587
+	2	2	0.304615	0.304615	0.370587
D	1	1	0.001923	0.001923	0.003410
D	1	2	0.001923	0.001923	0.003410
D	2	1	0.003077	0.003077	0.001590
D	2	2	0.003077	0.003077	0.001590

# of Iterations = 2

	#param	Ln(L)	Chi-square
H0: No Association	2	-539.16	0.00
H1: Markers Asso., Indep. of Disease	3	-539.16	0.00
H2: Markers and Disease Associated	6	-519.68	38.97

There are 8 Possible Haplotypes of These 3 Loci.
They are Listed Below, with their Estimated Frequencies:

Allele at Disease	Allele at Marker1	Allele at Marker2	Haplotype Frequency: Independent	Ind-Disease	w/Asso.
+	1	1	0.190385	0.190385	0.124384
+	1	2	0.190385	0.190385	0.124384
+	2	1	0.304615	0.304615	0.370616
+	2	2	0.304615	0.304615	0.370616
D	1	1	0.001923	0.001923	0.003411
D	1	2	0.001923	0.001923	0.003411
D	2	1	0.003077	0.003077	0.001589
D	2	2	0.003077	0.003077	0.001589

# of Iterations = 3

	#param	Ln(L)	Chi-square
H0: No Association	2	-539.16	0.00
H1: Markers Asso., Indep. of Disease	3	-539.16	0.00
H2: Markers and Disease Associated	6	-519.67	38.97

Methodology

Assumptions and observations

The disease has two alleles, a disease allele, *D*, and a normal allele, *d*, where the disease allele frequency, $p = P(D)$, is assumed to be known. There are two phenotypes, A = affected, and U = unaffected. Penetrances of being affected given genotypes DD, Dd, and dd, are f_{DD}, f_{Dd}, and f_{dd}, respectively, and are taken to be known. This general parametrization allows for dominant or recessive inheritance, possibly with phenocopies. The penetrances are assumed fixed.

Two sampling schemes are considered, (1) random sampling from the population, and (2) sampling conditional on an individual's disease phenotype. In the latter case, there are two samples of unrelated individuals, one of size n_A consisting of *A* phenotypes, and one of size n_U consisting of *U* phenotypes. On each individual, genotypes at a number of marker loci are observed, where there may be any number of alleles at the different marker loci. The outline in the next paragraph assumes complete marker typing; missing marker genotypes is not currently allowed.

Likelihood calculations

Unconditional case. The likelihood for the i-th affected or unaffected individual is given by $P(A_i, M_i)$ or $P(U_i, M_i)$, respectively, and the total likelihood over all data is simply $\Pi_i P(A_i, M_i)\Pi_j P(U_j, M_j)$, where *M* stands for an individual's phenotype at all marker loci. For simplicity, the outline below focuses on individuals of the *A* phenotype; calculations for *U* phenotypes are analogous to those for *A* phenotypes.

The likelihood for an individual may be calculated as $P(A, M) = \Sigma_g P(A, M|g)P(g)$, where *g* is a multilocus genotype (at all loci) consisting of two haplotypes, and the sum is taken over all possible genotypes. The number of these genotypes is $k(k + 1)/2$, with *k* being the number of multilocus haplotypes. The term $P(A, M|g)$ is given by the product of all penetrances at the different loci. At the marker loci, these penetrances are generally 0 or 1, but at the disease locus the penetrances are given by the three *f* values discussed above.

Unconditional probabilities of genotypes, *g*, are computed from the population haplotype frequencies in the usual manner.

Conditional case. When individuals are ascertained in two independent samples, conditional on their disease phenotype, the associated conditional likelihood (e.g., for an affected individual) is given by $P(A,M|A) = P(M|A) = \Sigma_g P(M|g,A)P(g|A)$. Rather than deriving explicit formulas for these expressions, it is easier to describe how to obtain them numerically. The following table below displays genotypes at the disease and one marker locus with alleles *B* and *b*:

	BB	Bb	bb	Sum
DD	...	...	...	p^2
Dd	...	...	...	$2p(1-p)$
dd	...	...	...	$(1-p)^2$

Each cell in the above table has an unconditional cell (genotype) probability, for example, P(g) = P(Dd, bb), which is given by the population haplotype frequencies. Based on the disease genotype in a cell, that

cell gives rise to an affected and unaffected individual with corresponding conditional probabilities, P(A|g) and P(U|g), where P(A|g) + P(U|g) = 1 and P(A) and P(U) are given by the penetrances at the disease locus. Thus, each cell contributes a term, P(A|g)P(g), to the total probability, P(A), of being affected, and a term, P(U|g)P(g), to the total probability, P(U), of being unaffected. In the sample containing *A* phenotypes, the sum over all cell probabilities must be equal to 1 rather than P(A); that is, for the affected sample, a cell probability is obtained as P(A|g)P(g)/P(A); for the unaffected sample, the corresponding cell probability is P(U|g)P(g)/P(U).

Gene counting

The method of gene counting is a particular form of the EM algorithm, which iteratively furnishes ML estimates of parameters. Consider, for example, the cell with genotypes *Dd* and *bb* in the above table. An individual in this cell contains a *Db* and a *db* haplotype. Therefore, these two haplotypes can be directly observed and included in a count of haplotypes. On the other hand, an individidual with genotypes *Dd* and *Bb* has one or the other of the phases DB/db and Db/dB. Each phase occurs with a probability given by the haplotype frequencies. According to the principle of gene counting, such a doubly heterozygous individual may be split into two components in the proportions given by the phase probabilities. With a given phase, haplotypes can again be counted but these counts contribute towards the total only with a weight (of less than 1) as given by the appropriate phase probability. This method is directly applicable to the unconditional case mentioned above.

In the case of the two phenotype samples, haplotypes must by design be counted separately in each sample. In a given sample, it is not only in multiply heterozygous individuals that haplotypes cannot be counted directly. Disease genotypes cannot generally be recognized either. For example, consider an affected individual with marker genotype *BB*. With general penetrances, this individual can have any one of the three disease genotypes, *DD*, *Dd*, or *dd*. Each of these three cases has probability of occurrence given by the joint genotypes, DB/DB, DB/dB, and dB/dB, which, in turn, are given by the haplotype frequencies. Thus, our affected *BB* individual is split into three components with associated relative probabilities (weights), which furnishes weighted haplotype counts.

With a given set of initial haplotype frequencies, the gene counting method furnishes in each phenotype sample a new count of haplotypes, which is converted into relative frequencies. For the sample of *A* phenotypes, this leads to a probability density, P(h|A), of haplotype frequencies, where *h* stands for any haplotype. Analogously, the sample of *U* phenotypes furnishes P(h|U). Since *A* phenotypes occur in the population with a frequency of P(A) and *U* phenotypes with a frequency of P(U), the population haplotype frequencies are given by P(h) = P(h|A)P(A) + P(h|U)P(U). These updated haplotype frequencies then replace the initial haplotype frequencies, and a new iteration is started. After a sufficient number of such iterations, the haplotype frequencies will be close to their MLE's.

Acknowledgment

Development of this program was partly supported by US Army Medical Research Acquisition Activity under award # DAMD17-94-J-4406. Any opinions, findings and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the view of the US Army Medical Research Acquisition Activity.

References

Xie X, Ott J (1993) Testing linkage disequilibrium between a disease gene and marker loci. *Am J Hum Genet* **53**, 1107 (abstract)

Terwilliger J, Ott J (1994) *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore